Big Data Systems and Techniques - Homework for June 2020

The scope of the homework is to train a classifier in a big data environment for determining the category (i.e. men's shoes) from a large dataset of categorized (labeled) products and their descriptions and apply the classifier on streams of twitter tweets about shopping offers.

The homework will be executed on the cluster we have installed in the class. To pass the class the minimum requirement is to have working the 3 node cluster in your google account with everything installed on Courses 1 to 8. Otherwise it is an automatic failure. So if you are not at this point - install the remaining components before start working in the homework tasks.

Task 1 (5% - 15% if you use Google Cloud SQL) - Get the data

In the downloads section of the class wiki there is a file called product.sql.zip. Download the file in your s01 server.

This is a postgres dump of database containing only 1 table temp_products with rows like this:

```
products=# select * from temp_products limit 1;

 product_id |            name             | upc_id |
descr
|                                  vendor_catalog_url
|                                      buy_url
| manufacturer_name | sale_price | retail_price |
manufacturer_part_no | country | vendor_id | category_name |
category_code | category_id
------------+------------------------------+--------
+-------------------------------------------------------------------
-------------------------------------------------------------------
-------------------------------------------------------------------
+-------------------------------------------------------------------
----------------------
+-------------------------------------------------------------------
------------------+------------------+------------+--------------
+----------------------+---------+-----------+---------------
+---------------+-------------
     313961 | Nili Lotan Wool Cocoon Coat |        | Nili Lotan
black fuzzy wool cocoon coat. Collarless, horizontal-stitch detail
at front, side-seam pockets. Concealed button closure at front.
Lined. Available in Black. Wool. Dry clean. Made in U.S. . | http://
www.shopstyle.com/p/nili-lotan-wool-cocoon-coat/459580123?
pid=uid8576-26123524-6 | http://www.shopstyle.com/action/
apiVisitRetailer?id=459580123&pid=uid8576-26123524-6 | Nili Lotan
Coats  |       1195 |         1195 |                      |
|           | Wool Coats    | wool-coats    |        1445
```

Research how you will insert this file in postgres db installed in your system and insert the data. If you like research postgres in Google Cloud SQL and use a managed instance to restore data there (this will give 10% more score).

If all correctly done you expect to have a DB with:

```
products=# select count(*) from temp_products;
  count
---------
 2080734
(1 row)
```

As deliverable describe the commands used to insert the data in postgres.

Task 2 - Create a parquet file (10%)

Locate the categories of shoes and a create a subset of data with shoes only.

Write a spark program that connects to postgresdb reads the data in DataFrame and writes the DataFrame in HDFS in parquet format.

As deliverable give the above program, the printSchema of the DataFrame and the *hdfs dfs ls* of the directory with the parquet files.

Task 3 - ML (15% data processing, 25% algorithm training)

Research Spark Documentation and select two classification algorithms.
The goal is to train classifiers to detect the category of each product using any information on the other columns (i.e. name, description, brand) except `category_name, category_code, category_id`.

Think what parameters you should tune. Select as K whatever you think best for the problem.

Write a spark program that reads the Task 2 parquet file in a DataFrame and processes the data in order to be suitable to be input in the algorithms.

Apply cross validation verbose to train the algorithm and tune the params. Select the params with maximum performance.

Save the best model to HDFS.

Run the training program in spark in distributed mode.

As deliverable give the execution commands, the above program, dumps of its execution in cross validation verbose and the model file.

Task 4 - Kafka (15%)

Write a Kafka producer that consumes tweets from twitter and posts them in a topic called “offers”. Subscribe in twitter to get your own developer credentials in order to access twitter APIs without problems. You should retrieve from tweeter topics searching with “*shopping offers shoes*” keywords. Try alternative searches to see what brings descriptions similar to the ones in db.

As deliverable give the above program and a dump of the tweets with timestamps proving that the program is working.

Task 5 - Spark streaming (25%)

If you did not manage to write the Kafka producer use console producer to continue.

Write a Spark program that:

Loads the best model saved on Task 3.

Consumes from Kafka topic *offers* in with Spark Streaming.

Performs the processing needed in tweets (like in Task 3) to be able to run the model evaluation.

Evaluates the tweets and writes the tweets and their evaluation in a parquet file and in a text file.

As deliverable give the above program, dumps of its execution printing tweets and categorization found, the printSchema of the DataFrame for eport and the *hdfs dfs ls* of the directory with the parquet files and some text files of its execution.